BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 10: e183035
doi: 10.3897/biss.10.183035

OPEN ACCESS

Conference Abstract

Fine-Tuning Language Models for Structured Botanical Trait Extraction and Species Comparison

Ayoub Nainia‡, Régine Vignes Lebbe‡, Jihad Zahir§

‡ Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France
§ UMMISCO, IRD, Computer Science Department, LISI, CADI Ayyad University, Faculty of Sciences, Marrakech, Morocco, Marrakech, Morocco

Corresponding author: Ayoub Nainia (ayoub.nainia@sorbonne-universite.fr)

Received: 19 Dec 2025 | Published: 16 Jan 2026

Citation: Nainia A, Vignes Lebbe R, Zahir J (2026) Fine-Tuning Language Models for Structured Botanical Trait Extraction and Species Comparison. Biodiversity Information Science and Standards 10: e183035.
https://doi.org/10.3897/biss.10.183035

Abstract

As artificial intelligence (AI) becomes central to biodiversity research, transforming morphological descriptions of species into structured data remains a challenge, particularly in botany, where species descriptions are rich in detail but lack formal structure. Despite the progress brought by transformer architectures (Vaswani et al. 2017) and their multilingual variants such as CamemBERT (Martin et al. 2020), most large language models (LLMs) are trained on general-domain English corpora, limiting their effectiveness for non-English biodiversity applications.

We present a two-stage information extraction pipeline that fine-tunes LLaMA (Large Language Model Meta AI)- based LLMs (Touvron et al. 2023) to extract botanical traits from French floristic texts and enable species-level comparison at scale. Trained on a curated corpus of legacy botanical descriptions and expert-validated French question-answer pairs, the system demonstrates how domain-specific fine-tuning can bridge the gap between narrative botanical knowledge and structured trait data (Fig. 1). At the user-interface layer, we add a third stage that automatically converts the trait-value pairs generated by the question-answering module into a structured CSV table for downstream use.

© Nainia A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

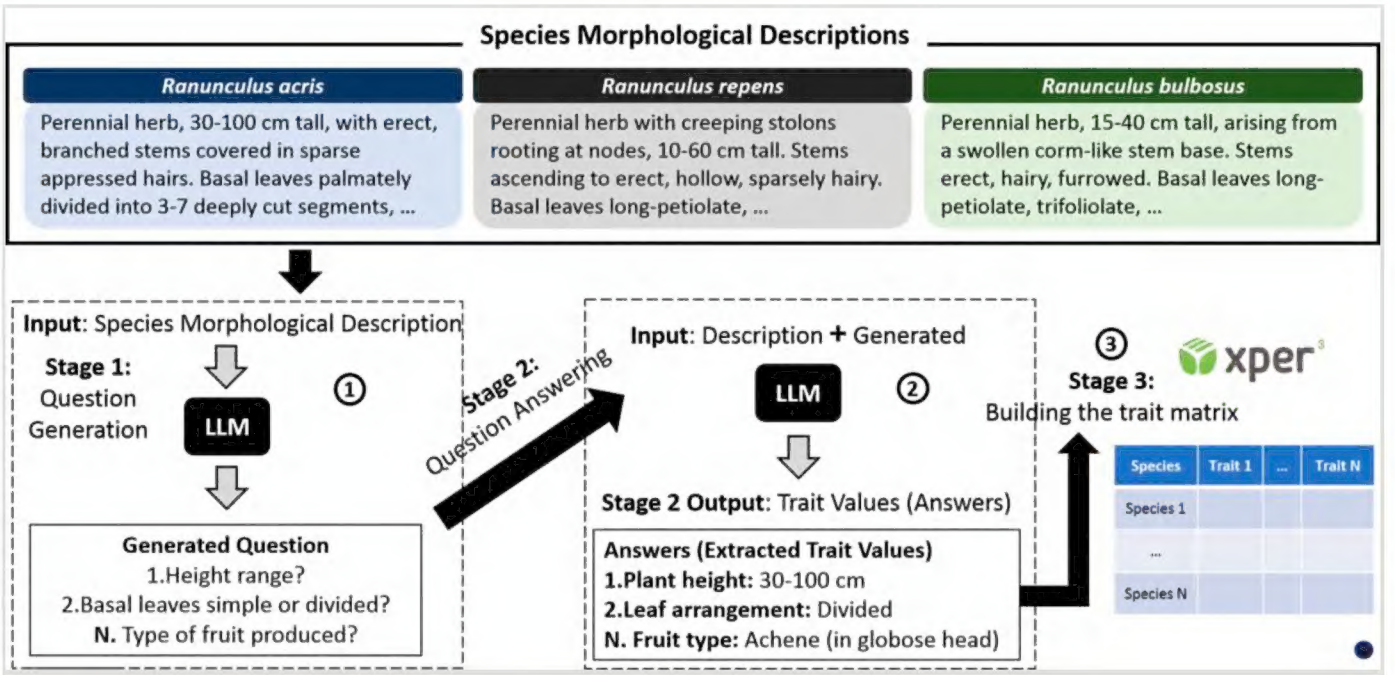
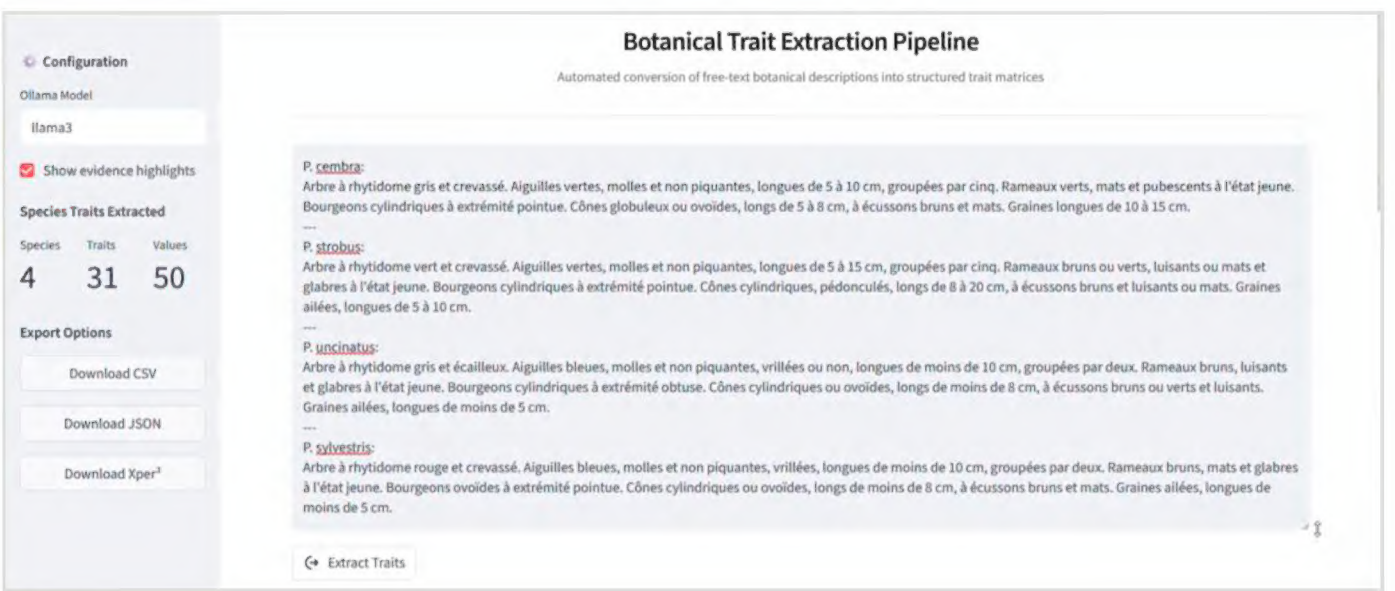

Species Morphological Descriptions
Ranunculus acris
Perennial herb, 30-100 cm tall, with erect, branched stems covered in sparse appressed hairs. Basal leaves palmately divided into 3-7 deeply cut segments, ...
Ranunculus repens
Perennial herb with creeping stolons rooting at nodes, 10-60 cm tall. Stems ascending to erect, hollow, sparsely hairy. Basal leaves long-petiolate, ...
Ranunculus bulbosus
Perennial herb, 15-40 cm tall, arising from a swollen corm-like stem base. Stems erect, hairy, furrowed. Basal leaves long-petiolate, trifoliolate, ...
Input: Species Morphological Description
Stage 1: Question Generation
LLM
①
Generated Question
1.Height range?
2.Basal leaves simple or divided?
N. Type of fruit produced?
Stage 2: Question Answering
Input: Description + Generated
LLM
②
Stage 2 Output: Trait Values (Answers)
Answers (Extracted Trait Values)
1.Plant height: 30-100 cm
2.Leaf arrangement: Divided
N. Fruit type: Achene (in globose head)
③
Stage 3: Building the trait matrix
xper
Species
Trait 1
...
Trait N
Species 1
...
Species N

Figure 1.

Two-stage LLM pipeline for botanical trait extraction and comparison. Credit: A. Nainia. License: CC BY 4.0.

Our approach unifies question generation and trait extraction within a single information extraction workflow. First, predefined botanical traits (e.g., leaf shape, flower color) are reformulated as natural-language questions tailored to each species description. Then, using these questions as prompts, the model identifies and extracts the corresponding trait values: for example, extracting "lanceolate" as the standardized leaf shape from "lanceolate leaves." This converts unstructured text into consistent character-state pairs aligned with botanical ontologies. A proof-of-concept showing the input species descriptions processed by the pipeline is shown in Fig. 2.


Configuration
Ollama Model
llama3
Show evidence highlights
Species Traits Extracted
Species
Traits
Values
4
31
50
Export Options
Download CSV
Download JSON
Download Xper³
Botanical Trait Extraction Pipeline
Automated conversion of free-text botanical descriptions into structured trait matrices
P. cembra:
Arbre à rhytidome gris et crevassé. Aiguilles vertes, molles et non piquantes, longues de 5 à 10 cm, groupées par cinq. Rameaux verts, mats et pubescents à l'état jeune. Bourgeons cylindriques à extrémité pointue. Cônes globuleux ou ovoïdes, longs de 5 à 8 cm, à écussons bruns et mats. Graines longues de 10 à 15 cm.

P. strobus:
Arbre à rhytidome vert et crevassé. Aiguilles vertes, molles et non piquantes, longues de 5 à 15 cm, groupées par cinq. Rameaux bruns ou verts, luisants ou mats et glabres à l'état jeune. Bourgeons cylindriques à extrémité pointue. Cônes cylindriques, pédonculés, longs de 8 à 20 cm, à écussons bruns et luisants ou mats. Graines ailées, longues de 5 à 10 cm.

P. uncinatus:
Arbre à rhytidome gris et écailleux. Aiguilles bleues, molles et non piquantes, vrillées ou non, longues de moins de 10 cm, groupées par deux. Rameaux bruns, luisants et glabres à l'état jeune. Bourgeons cylindriques à extrémité obtuse. Cônes cylindriques ou ovoïdes, longs de moins de 8 cm, à écussons bruns ou verts et luisants. Graines ailées, longues de moins de 5 cm.

P. sylvestris:
Arbre à rhytidome rouge et crevassé. Aiguilles bleues, molles et non piquantes, vrillées, longues de moins de 10 cm, groupées par deux. Rameaux bruns, mats et glabres à l'état jeune. Bourgeons ovoïdes à extrémité pointue. Cônes cylindriques ou ovoïdes, longs de moins de 8 cm, à écussons bruns et mats. Graines ailées, longues de moins de 5 cm.
Extract Traits

Figure 2.

Overview of French species descriptions processed by the interface, with extracted botanical traits highlighted. Credit: A. Nainia (CC BY 4.0).

The LLaMA model was fine-tuned via Low-Rank Adaptation (LoRA) (Hu et al. 2021) on ~17,000 manually verified French question-answer pairs derived from floristic

descriptions across multiple French-language flora, including: *Flore de Nouvelle-Calédonie* and *Flore du Cameroun*. Building on FloraNER (Named Entity Recognition) (Nainia et al. 2024a) and earlier CamemBERT-based pipelines (Nainia et al. 2024b), the current system, F-LoRA-QA (Nainia et al. 2025b), substantially improves structured trait generation from domain-specific descriptions.

In evaluation, F-LoRA-QA outperformed an untuned LLaMA baseline with ~4× higher BLEU score (Bilingual Evaluation Understudy), a 16% gain in semantic similarity, and exact-match accuracy rising from 2% to 24%. Expert botanists further validated trait accuracy, completeness, fluency, and terminology, supporting critiques of over-reliance on automatic metrics for domain question answering (Nainia et al. 2025a).

A key application is the enrichment of descriptor-based systems like Xper³ (Kerner et al. 2025), which rely on structured trait matrices to describe and differentiate taxa. Fig. 3 and Fig. 4 illustrate trait discovery, value extraction, and a comparison matrix with evidence links, supporting descriptor systems and interoperable biodiversity databases.

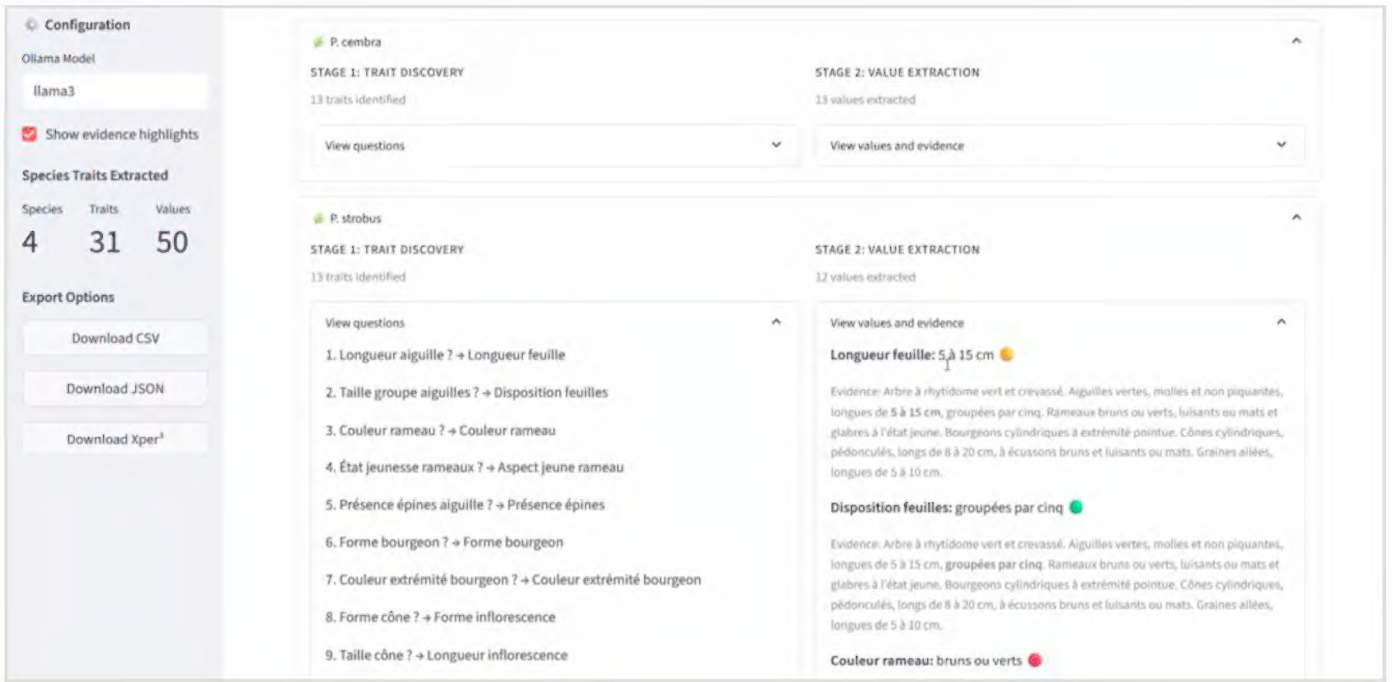

Configuration
Ollama Model
llama3
Show evidence highlights
Species Traits Extracted
Species
Traits
Values
4
31
50
Export Options
Download CSV
Download JSON
Download Xper³
P. cembra
STAGE 1: TRAIT DISCOVERY
13 traits identified
STAGE 2: VALUE EXTRACTION
13 values extracted
View questions
View values and evidence
P. strobus
STAGE 1: TRAIT DISCOVERY
13 traits identified
STAGE 2: VALUE EXTRACTION
12 values extracted
View questions
View values and evidence
1. Longueur aiguille ? → Longueur feuille
2. Taille groupe aiguilles ? → Disposition feuilles
3. Couleur rameau ? → Couleur rameau
4. État jeunesse rameaux ? → Aspect jeune rameau
5. Présence épines aiguille ? → Présence épines
6. Forme bourgeon ? → Forme bourgeon
7. Couleur extrémité bourgeon ? → Couleur extrémité bourgeon
8. Forme cône ? → Forme inflorescence
9. Taille cône ? → Longueur inflorescence
Longueur feuille: 5 à 15 cm
Evidence: Arbre à rhytidome vert et crevassé. Aiguilles vertes, molles et non piquantes, longues de 5 à 15 cm, groupées par cinq. Rameaux bruns ou verts, luisants ou mats et glabres à l'état jeune. Bourgeons cylindriques à extrémité pointue. Cônes cylindriques, pédonculés, longs de 8 à 20 cm, à écussons bruns et luisants ou mats. Graines ailées, longues de 5 à 10 cm.
Disposition feuilles: groupées par cinq
Evidence: Arbre à rhytidome vert et crevassé. Aiguilles vertes, molles et non piquantes, longues de 5 à 15 cm, groupées par cinq. Rameaux bruns ou verts, luisants ou mats et glabres à l'état jeune. Bourgeons cylindriques à extrémité pointue. Cônes cylindriques, pédonculés, longs de 8 à 20 cm, à écussons bruns et luisants ou mats. Graines ailées, longues de 5 à 10 cm.
Couleur rameau: bruns ou verts

Figure 3.

Trait discovery (Stage 1) and value extraction (Stage 2) for *Pinus strobus*, with answers linked to source species descriptions. Credit: A. Nainia (CC BY 4.0).

Keywords

question answering, natural language processing, large language models, information extraction, computational biodiversity

Presenting author

Ayoub Nainia

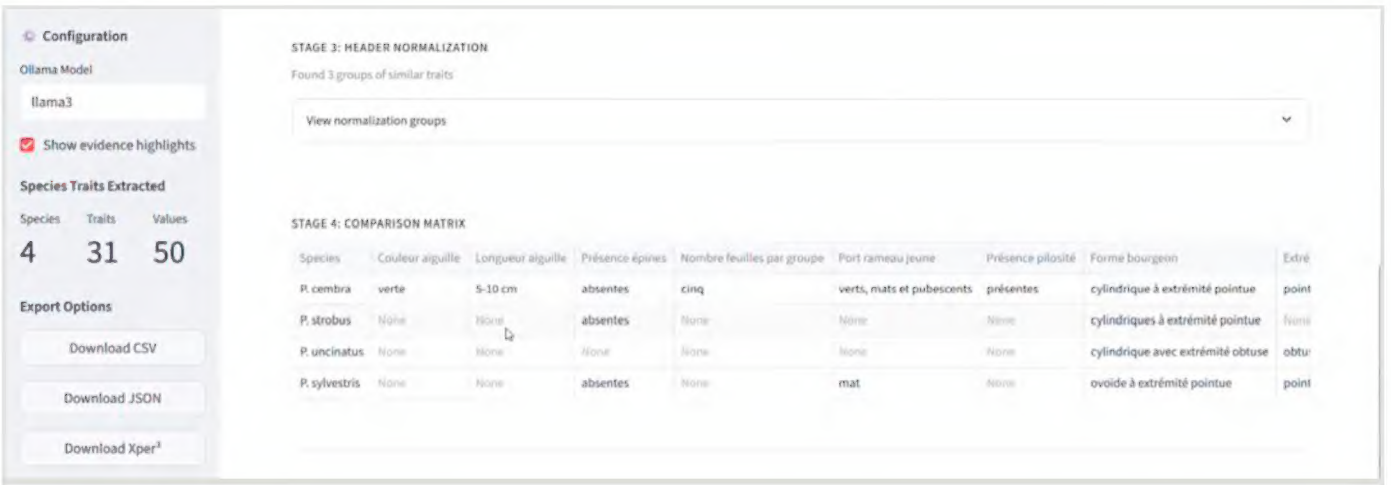


Figure 4.

Species comparison matrix provided by the pipeline, showing normalized trait values across species. Credit: A. Nainia. License: CC BY 4.0.

Presented at

Living Data 2025

Acknowledgements

This project was co-funded by the EU Horizon Europe programme (MSCA Cofund SOUND.AI, Grant No. 101081674) and the e-COL+ project (ANR-21-ESRE-0053).

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) LoRA: Low-Rank Adaptation of Large Language Models. URL: https://doi.org/10.48550/arXiv.2106.09685
- Kerner A, Saliba EM, Bouquin S, Portier R, Vignes-Lebbe R, et al. (2025) An Xper3 reference guide for taxonomists: a collaborative system for identification keys and descriptive data. European Journal of Taxonomy 987: 281-302. https://doi.org/10.5852/ejt.2025.987.2875
- Martin L, Muller B, Ortiz Suárez PJ, Dupont Y, Romary L, de la Clergerie É, Seddah D, Sagot B (2020) CamemBERT: a Tasty French Language Model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics https://doi.org/10.18653/v1/2020.acl-main.645
- Nainia A, Vignes-Lebbe R, Chenin E, Sahraoui M, Mousannif H, Zahir J, et al. (2024a) FloraNER: A new dataset for species and morphological terms named entity recognition in French botanical text. Data in Brief 56 URL: https://doi.org/10.1016/j.dib.2024.110824

- Nainia A, Vignes-Lebbe R, Chenin E, Sahraoui M, Mousannif H, Zahir J, et al. (2024b) A Transformer-Based Nlp Pipeline for Enhanced Extraction of Botanical Information using Camembert on French Literature. NLP & Information Retrieval59-78. https://doi.org/10.5121/csit.2024.140605
- Nainia A, Vignes-Lebbe R, Mousannif H, Zahir J, et al. (2025a) Beyond BLEU: Ethical Risks of Misleading Evaluation in Domain-Specific QA with LLMs. In: Picazo-Izquierdo A, Estevanell-Valladares EL, Mitkov R, Guillena RM, Cerda RG (Eds) Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models. RANLP 2025, Bulgaria. [ISBN 978-954-452-102-8]. https://doi.org/10.26615/978-954-452-102-8_009
- Nainia A, Vignes-Lebbe R, Mousannif H, Zahir J (2025b) F-LoRA-QA: Finetuning LLaMA Models with Low-Rank Adaptation for French Botanical Question Generation and Answering. In: Angelova G, Kunilovskaya M, Escribe M, Mitkov R (Eds) Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era. RANLP 2025, Bulgaria. 787--796 pp. [ISBN 978-954-452-098-4]. https://doi.org/10.26615/978-954-452-098-4-091
- Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023) LLaMA: Open and Efficient Foundation Language Models. URL: https://arxiv.org/abs/2302.13971
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I (2017) Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762